DRAFT: STILL AWAITING FINAL COMMENT FROM PARTICIPANTS

Report of the Workshop to design simulation-based performance tests for evaluating methods used to infer population structure from genetic data

1 INTRODUCTORY ITEMS

The meeting was held at the the Center for Marine Biodiversity and Conservation (CMBC), Scripps Oceanographic Institute, La Jolla, from 21 to 24 January 2003. Day 1 consisted of presentations on topics related to whales, management, genetics and stock structure. Discussions took place on days 2-4. A list of Participants is given as Annex A.

Donovan chaired the meeting. On behalf of the participants, he thanked the Steering Group and particularly Karen Martien for organizing the meeting and CMBC for providing a venue. Best, Bravington, Hoelzel, Perrin and Taylor acted as rapporteurs.

Annex B shows the adopted agenda. The organisation of the sections in this report has been changed somewhat for clarity, and the text draws both from background presentation and subsequent discussions. Documents available to the Workshop are listed in Annex C.

2 OBJECTIVES

Understanding population structure in the marine environment has proven the least tractable of the many uncertainties that vex management of wild living resources. Recent advances in the field of genetics gives the prospect of a powerful new tool to provide data on population structure. However, analytical tools have been almost exclusively developed to address evolutionary questions. When using these new tools to address management issues, questions arise as to how well genetic data and available genetic analytical methods perform at defining population structure at the level relevant to management.

Over the last two decades, there has been increasing consensus that the best way to deal with such uncertainties (like the problem of estimating population structure) is through the adoption of formal management procedures that have been extensively tested by computer simulation to ensure that they work well across a range of plausible scenarios that might represent the true situation. Acceptable procedures are those that meet agreed conservation and exploitation objectives to the extent possible, given the inevitable trade-off between them. The IWC has been at the forefront of this approach to management, through the development of its Revised Management Procedure (RMP) and more recently its Aboriginal Whaling Management Procedure(s).

The RMP includes an algorithm which sets the total catch limit across a geographical area based on historical catch data and a time-series of abundance estimates. There are also various options for setting area-specific catch limits on a smaller spatial scale, to guard against unwanted local over-depletion. The catch limit algorithm applies to all stocks and does not need to be revisited in individual cases, but choices about area-specific splits have to be made anew for each stock where whaling is contemplated. However, there is no formal mechanism for deciding how many boundaries between areas there should be, or where they should go. In practice, the IWC has developed an initial set of *Small Areas* in a somewhat *ad hoc* fashion. Without objective methods for using genetic stock structure data in selecting management areas, scientific consensus can be difficult to obtain, and case-specific testing becomes an inordinately complex and slow process. As the IWC moves forward on management rules for aboriginal whaling--- and as other management bodies grapple with issues such as sustainability of spatially-concentrated by-catch--- the need for a more systematic approach grows ever greater.

The quality and quantity of genetic data have increased over recent years, and several new methods have been added to traditional hypothesis testing approaches to illuminate population structure. However, most of these methods have not been subjected to simulation performance testing, making comparative performance judgments impossible. To check how well such methods would perform in real management contexts, simulation tests are required. This led to the IWC's recommendation to hold the present workshop to focus on designing simulation-based performance tests for evaluating methods used to infer population structure from genetic data. The goals of the workshop are as follows:

(1) Outline a set of simulations that capture the types of population structure likely to be encountered in migratory whales

(2) Outline a set of realistic sampling schemes that reflect the types and sizes of samples likely to be available for migratory whale species (e.g., sampling on feeding grounds, breeding grounds, migratory routes, or combinations thereof)

(3) Design performance measures that reflect the likely performance of the analytical methods with respect to management under both the RMP and the AWMP.

(4) Specify format of simulated datasets to be made available to developers.

Thus the workshop will try to develop a "road map" for simulation testing of boundary placement methods, and to elaborate some of the details. In addition it will help to publicise the problem of how to set management boundaries using genetic data, and to stimulate the interest of researchers outside the IWC in developing practical methods. Setting management boundaries is a critical issue not just for direct hunting (e.g. commercial and aboriginal whaling, fisheries) but also for dealing with incidental mortality (e.g. by-catch of small cetaceans or birds); in fact it is essential for the practical conservation of widely-distributed populations of any mobile species.

3 BACKGROUND

3.1 Simulation testing and management procedures in general

Regardless of the aspect of management procedure being tested--- e.g. quota setting rules, boundary placement methods, (in fisheries) effort limits, time/area closures--- the basic procedure is the same. The simulation framework is established in isolation from the management procedures to be tested. The framework consists of a set of "operating models" that describe in detail the relevant aspects of (and span the range of uncertainties about):

(i) population biology (e.g. how animals move between areas; how fast they reproduce and die naturally);

(ii) any details of exploitation not explicitly covered by management (e.g. exactly where catches will be taken if no boundaries are enforced);

(iii) data generating mechanisms (e.g. how often will abundance estimates be made, and how much noise or bias will they have? how many genetic samples will be taken, how many loci will be examined, and what will the error rate be?).

Different assumptions can be accommodated by varying the relevant operating model. Any particular combination of assumptions constitutes a "scenario".

Testing a procedure to see if it meets its stated objectives obviously requires the specification of performance statistics: related for example to catch, population size, proportion of alleles lost from the population etc. Such performance statistics may depend on quantities that stay hidden within the simulation, and which are not passed on explicitly to the management procedure being tested (e.g. the true rather than estimated population size).

The various management procedures to be tested can have very different algorithms or assumptions built in to them, but should exhibit certain features: (a) be able to cope with data in the format provided by the data generation step of the simulation (although it does not have to use all the data); (b) not depend on data that the simulation won't provide (e.g. population growth rate); (c) not require human intervention; and (d) deliver its result in a specified form (e.g. "the catch limit this year in area 1 is X"). A management procedure comprises several components: e.g. one to set boundaries; one to calculate catch limits. It is not necessary for each component to be applied each year. For example, the boundary-setting method might be applied only once at the start of the simulation, while the catch rule might be applied every 5 years. These timing issues are usually thought of as fixed features of the simulation framework that will be kept constant for all methodss tested, rather than adjustable properties of any one method.

To test one particular management procedure under one particular scenario, the *modus operandi* is this:

1. Simulate some historical data, according to the scenario.
2. Apply the management procedure to those data, to determine what management action will happen in the first simulation-year.
3. Simulate the population dynamics over the simulation-year, according to the management action and the biology and exploitation details in the scenario.
4. Compute performance statistics for the simulation-year.
5. Generate any data needed by the management procedure (e.g. an abundance estimate).
6. Move on a year; go back to step 2 until 100 simulation-years (or however long) have passed.
7. Summarize the performance statistics.

Each candidate management procedure gets tested in each scenario, so that the performance of different management procedures can be compared under similar conditions. Usually, two different computer programs are used; one ‘common control program’ which handles steps 1 and 3-7, and a set of "slave" programs, one per management

procedure, to carry out step 2 when specified. To test a new management procedure, all that is needed is a new "slave" program.

3.2 Issues specific to boundary placement in the IWC

A boundary placement algorithm is thus one part of a management procedure; there also needs to be a component that sets catch limits. In the IWC context, this latter part is already specified, so that from a simulation perspective, it is only the boundary placement method that needs to be considered (and programmed)[1]. The catch control law essentially represents a feedback loop that adjusts quotas up or down if the time-series of abundance estimates indicates that productivity is higher or lower than initially assumed. Given the objectives established by the IWC, the catch control law for the RMP is highly conservative. The catch limits are set to zero if there is a possibility that the stock might be at less than 72% of carrying capacity. However, the actual level of depletion may be substantially under- or over-estimated if stocks are mis-defined.

The perceived success of any method will depend on the performance criteria. In particular, it will depend on the choice of "unit-to-conserve". The RMP has an agreed maximum overall level of depletion for a 'stock' that is tolerable but does not include a clear definition of what a 'stock' is. In a simple sense, choosing a 'unit-to-conserve' is equivalent to choosing the spatial scale at which over-depletion is to be avoided. There are as yet, no management guidelines for unit-to-conserve, and indeed no single overall quantitative objective for management. Thus the scientific task is to design a set of performance measures that report results succinctly but in a way that allows depletion to be assessed on different spatial scales.

A further question to ask of any boundary placement method is how often should it be reconsidered over the simulation period (c.f. new population estimates become available every *x* years). As whaling (or by-catch) proceeds, more genetic data will accumulate, presumably leading to better information about population structure. It is possible to envisage a feedback boundary placement method that adjusts boundaries on the basis of new data. In fact, the present IWC RMP and AWMP have provisions for this via the "*Implementation Review*" system. However, it was **agreed** that at this stage it will be challenging enough to develop methods without feedback. However, such feedback methods will have to be considered in order to ensure relevance to IWC practice.

The 'interesting' situations from the IWC's point of view concern limited dispersal between stocks. Specifically, this means cases where dispersal is fast enough to leave only a small amount of genetic differentiation, but too slow to allow "rapid" repopulation of a depleted area from neighbouring areas. The systematic use of genetic data for management boundary selection in such cases is a new approach. It was **agreed** that the initial focus of testing will therefore be on exploring how different genetic approaches perform and on how to improve promising approaches, rather than with any view to choose the best method.

Given this exploratory focus, it was **agreed** that the simulation scenarios should be designed to identify clearly differences between the performances of various boundary placement methods. The scenarios need to reflect general features of whaling management, but should not focus too heavily on individual cases. This will allow clear comparisons to be made, and help to ensure the general relevance of any conclusions reached.

The development of a simulation framework, as well as the development of management methods, is an iterative process. The scenarios considered will initially be few in number and fairly simple. As methods evolve, more realistic complexity will be added to the scenarios, and some features may be dropped because results show that they make little difference. The same holds for performance measures. The lesson from the IWC's lengthy RMP and AWMP development process is that scientific insight will be only gained by using a stage-by-stage process that allows time for reflection and interaction among the participants.

More details on these points can be found later in the report.

3.3 Population biology of whales

Annex D presents overviews of whale demography, life history, migratory patterns and population structures, concentrating on the large baleen whales which are the main focus of the IWC. These whales in general are long-lived, with average life expectancies ranging from 16 to over 100 years, and have low reproductive rates. Maximum population growth rates are therefore low (of the order of 3-10% per year), and have historically been surpassed by annual fishing rates, resulting in declines to a few percent of original abundance in several cases.

Most baleen whales migrate seasonally over long distances between winter calving and breeding grounds in subtropical/tropical waters and polar feeding grounds in summer. One of the best-studied species is the humpback

[1] Actually, there are several versions of the quota-setting rule, using different ways of splitting quota across pre-defined areas; see section **. Each boundary placement algorithm could be tested several times, in combination with the different versions of quota-setting.

whale. Large concentrations of breeding animals tend to occur in shallow waters near coasts or island groups. Animals from one breeding ground can migrate to different feeding grounds. In the North Atlantic, strong fidelity of animals to a feeding ground is observed. This is presumably maintained through the calf accompanying the mother to the feeding ground during the 6-9 month lactation period. The information for other baleen whales (apart from the gray whale) is much less, and the location and nature (concentrated or dispersed) of the breeding grounds are frequently unknown. Some baleen whale migrations are often structured by sex, age and reproductive status, such that arrival and residence times on breeding and feeding grounds can differ substantially depending on these factors. It is possible that migration timing may also be systematically related to feeding or breeding ground, so that population substructuring may occur in time as well as in space.

Little is known about mating systems but it is likely that these systems differ quite dramatically among the large whales. For example, right and bowhead whales have extremely large testes for their body size and have been observed in mating groups consisting of many males to a single female, suggesting potential sperm competition. Such group mating behavior is not observed among the humpbacks, in which males sing, escort females and have much smaller testes per body size. The variance of male contribution to the next generation will effect the effective population size that determines the rate of genetic differentiation in nuclear DNA. Such differences in mating system will not effect effective population size for mtDNA, which depends on the number and variance in reproduction only among females.

3.4 Patterns of whaling

Depending on the behaviour and range of the species, whaling has historically taken place on feeding grounds, breeding grounds, and/or migration routes. Whaling on breeding grounds is no longer significant (e.g. a small catch of humpback whales is taken from the island of Bequia, St Vincent and The Grenadines) but it still occurs on the feeding grounds (e.g. catches of common minke whales in the northeastern Atlantic; Antarctic minke whales in the Antarctic; gray whales off Chukotka) and/or migration routes (e.g. common minke whales in the North Pacific; bowhead whales off Alaska and Chukotka; gray whales off Washington State). The IWC's RMP was designed for management of baleen whales caught on feeding grounds, but is now also being applied to cases where whaling may take place on migration routes as well as feeding grounds (e.g. common minke whales in the North Pacific off Japan). Whaling may not cover the whole range of a stock, and in the absence of regulations to the contrary, catches may not be evenly distributed across the range, because it is easier and cheaper to whale as close to home as practical.. Aboriginal subsistence whaling is always carried out close to shore. Commercial whaling can vary depending on the nature of the operation (e.g. coastal whaling where vessels are rarely at sea for more than 2 days versus pelagic whaling where vessels can remain at sea for months). Historic whaling grounds have sometimes been relatively small, which may result in limited and somewhat patchy data coverage.

3.5 Genetic data available for whales

The amount of genetic data available for whales is highly variable. Humpback whales are the best sampled from the perspective of sampling on both the breeding and feeding grounds. Both mtDNA and nuDNA markers have been used and yield different perspectives on population structure because the dynamics differ for the two different modes of inheritance. The species with the next highest sample size are the different minke whale species that are currently whaled. There are essentially no data from the breeding grounds for these species as the locations of these grounds remains unknown. Most other whales species have far <1% of the population sampled and for some species, such as the Sei whale, there are only a few samples in some ocean basins.

The most commonly used marker for population structure questions is sequences (300-900+ basepairs) of the mtDNA d-loop. To date this marker has proven the best at detecting population structure on average. Microsatellites are also used with most studies using few than 20 loci.

3.6 Techniques for simulating population genetics

There are three primary techniques employed in simulating population genetic data: Coalescent models, birth-death models and individual-based models (IBMs). In recent years, coalescent models have become the most popular of these techniques. Coalescent models generate a phylogeny for a simulated sample by using coalescent theory to project back in time to the most recent common ancestor of the simulated sample (see Hudson, **year**, for a review of coalescent theory). Genetic data are then generated by simulating forward in time on the phylogeny according to a pre-specified mutation model. Because only those individuals that actual contributed genetic material to the final sample are simulated, coalescent models are much faster than other methods (see below) of simulating genetic data. However, this computational simplicity comes at the cost of rigid, often biologically unrealistic, assumptions. All individuals are treated identically within the simulation, making it impossible to incorporate such complexities as sex-biased dispersal and pronounced polygyny.

Birth-death models involve projecting an entire population(s) forward through time and then drawing samples once the population(s) reaches mutation/migration/drift equilibrium. Consequently, birth-death models are considerably slower than coalescent models. However, they have greater flexibility for incorporating biological realism. Vital rates and behaviour are specified for specific classes of individuals (e.g., adult females, adult males and juveniles), which allows the model to incorporate features such as sex-biased dispersal, age-specific demography and moderate levels of polygyny/polyandry. Nonetheless, there is a limit to the level of biological realism that can be included in a birth-death model. For instance, because individual histories are not tracked, birth-death models cannot be used to generate relatedness-data or to simulate extreme polygyny, which requires male reproductive success to be correlated across years.

The simulation technique that employs the least number of assumptions is individual-based modelling (IBM), which can be custom designed to fully incorporate biological complexity. In an IBM, each individual is unique; the vital rates and behaviour of an individual can vary depending on the individual's spatial location, genetic makeup or individual history. The relationships between individuals can be tracked. Thus, nearly any level of biological realism can be accommodated. These models are not frequently used for genetic simulations because until recently computational limitations constrained their utility, i.e. these models tend to be much slower and require much more memory. Because every individual in the population is modelled separately, these computational constraints become pronounced for large abundances.

3.7 Methods for boundary placement

Before the workshop, the Steering Committee had identified eight methods that are currently available for using genetic data to evaluate population structure. These can potentially form the basis of boundary placing methods (note that most were not designed for boundary placement *per se*.) that could therefore be evaluated within a simulation framework. During the first day of the workshop, an expert in each method (usually its developer) gave a brief presentation describing the method. The methods discussed were: permutation-based hypothesis test with a null hypothesis of panmixia (Roff and Bentzen), Migrate (Beerli and Felsenstein), Structure (Pritchard et al.), Boundary Rank (Martien and Taylor), the method of Cui et al., Centroid (Archer), spatial auto-correlation (Tiedemann, Cassens et al.), and genetic mark-recapture (**ref**).

Table 1 provides a summary of the methods' data requirements, assumptions, original aim, and "output requirements" (i.e. what extras the method would need before it could be automated for management boundary setting in a simulation trial). Some traditional genetic descriptions of population structure (e.g. Moritz' ESU – evolutionary significant unit) are not likely to be of interest to management, because a single ESU could cover distinct parts of a population that in effect have completely separate population dynamics. Methods designed to pick up ESU-level distinctions, then, may not prove very effective for setting management boundaries. The question of the scale (resolution) of potential methods is important.

All of the methods need at least one extra feature before they can be automated for setting management boundaries in a simulation. Some kind of threshold value must always be chosen externally, e.g. for deciding where to stop subdividing areas, or at what level to reject a null hypothesis. This is not purely a statistical issue, but is linked to the requirements of management. For methods that explicitly estimate a migration rate, it is likely that the appropriate threshold for deciding to manage adjacent regions as a single unit rather than as two units, will be when the estimated migration rate between them exceeds some moderate fraction of the intrinsic rate of population growth (though other types of criterion can certainly be envisaged). It is at that level that migration becomes significant in population dynamics terms. An appropriate value for the fraction could be determined through experience in the simulation trials themselves. Unfortunately, migration rates of that magnitude are likely to be the most difficult to interpret. At very low migration rates, it is easy to detect a difference, and at higher migration rates, there is no need to manage separately.

As well as a "decision threshold", many methods require pre-specified boundaries across which hypotheses are to be tested or rates to be estimated. Other methods only generate hypothesized boundaries that still need to be evaluated before final stocks are defined. The workshop discussed several approaches to testing methods that only provide a partial solution to the problem of defining stocks, as outlined in section 6.3.

One approach in common use for studying population structure is Nested Clade Analysis (Templeton et al., 1995). Unfortunately, NCA is not viable for simulation testing because it requires a large number of decisions that need to be made "by hand" and cannot be easily automated (yet). Nevertheless, the questions that NCA addresses are certainly relevant to setting management boundaries, and automatic algorithms that attempt to tackle the same issues would certainly be worth testing. It was suggested that the creators of NCA be contacted to see if they are interested in developing a version of NCA that can be evaluated in an automated, simulation-testing framework.

There are other methods that deserve examination but were not discussed at the workshop; these include GenBar (Berthoud et al., in review), Spatial analysis of molecular variance (Dupanloup et al., 2002), Monmonnier's clustering

(Monmonnier, 1973) and "wombling" (Womble, 1951); there may be others. The simulation testing procedure should be structured to allow straightforward tests of other methods, e.g. by archiving simulated genetic data on the Internet, and providing an easy-to-use "common control program" that requires only a set of management boundaries as inputs.

4 SIMULATION DETAILS

4.1 Population biology

4.1.1 Archetypes

The Workshop **agreed** that the simulation exercise would need to be split into (at least) two phases. It recommends that in the first phase, five simple archetypes of population structure should be considered. For simplicity, this initial set should not include temporal components or sex-segregation (although these must be considered in the future). These are illustrated in Fig. 1 (from WP3)

Archetype I consists of a single panmictic population--- the "null model".

Archetype II is a simple stepping-stone model, with dispersal occurring only between adjacent stocks. Both a two-stock and a three-stock scenario will be used. This archetype was included for several reasons.

(1) Stepping-stone models have been extensively studied, which will facilitate comparison of our results to previous analytical and simulation studies.

(2) Their simplicity will aid in interpreting and validating the results of the simulations.

(3) The general applicability of stepping-stone models to many terrestrial as well as marine species will render the results of interest and utility to a broad conservation and management audience.

(4) Although migration is not explicitly incorporated, Archetype II is functionally identical to a migratory species in which there are separate breeding stocks, each of which migrates to a unique feeding area without crossing the migratory path of any other stock.

Archetype III is diffusion-type isolation-by-distance. In this model there are no distinct populations. Rather, density is even across the range and haplotype/allele frequencies change in a clinal fashion; genetic structuring is simply due to the fact that individuals only move a short distance over their lifetimes. There was some discussion concerning the relevance of a diffusion-type model to whale biology. A diffusion-type model implicitly assumes no effect of social structure, i.e. each individual acts independently of all other individuals. This assumption does not well capture a social mammal that may associate in “herds”. Nonetheless, it was **agreed** that this archetype would be included since it will present a significant challenge for boundary-setting methods; there are no discrete stocks, and yet it is important not to manage as a single unit. From a simulation perspective, a true diffusion process would require an entirely different simulation framework than the rest of the archetypes. However, this archetype can be modelled as an extension of Archetype II to a large number of discrete breeding stocks with limited dispersal that are sampled on a feeding ground where they have substantial overlap; this results in a continuous gradient in haplotype/allele frequencies.

Archetype IV consists of two migratory populations that have separate breeding grounds but whose feeding grounds overlap by a proportion X, where X is the proportion of each population that is in the zone of overlap. The Workshop **agreed** to examine initially two values for X: 0.5 and 1.0. The case where X = 1.0 is equivalent to two separate breeding stocks that share a single feeding ground. In this archetype, dispersal occurs when an animal changes breeding stocks.

Archetype V consists of a single breeding stock with two separate, non-overlapping feeding grounds. Animals follow their mothers to the feeding ground and subsequently exhibit strong feeding ground fidelity. Dispersal occurs between feeding stocks due to females occasionally changing feeding grounds. Archetype V is equivalent to panmixia (Archetype I) for nuclear markers, but not for mtDNA markers or for individual genetic tagging data.

The stock separation in all the above archetypes is along one dimension, which could represent longitude, latitude, or time. These archetypes therefore effectively cover cases such as a set of substocks which migrate past a single whaling point at different times. The words "feeding grounds" could equally well be replaced by "migratory corridor" in all the archetypes; animals that are heading for different feeding grounds might go past the sampling or whaling station at different times.

These archetypes are mainly geared towards baleen whales rather than odontocetes. At present there are no plans within the IWC to develop management procedures for sperm whales.

In the context of bycatches and direct catches, management of odontocetes is important but it would be difficult to produce a generalized archetype of population structure given their complex and highly variable social systems.

Consideration should be given to developing an odontocete archetype in future, though, with particular emphasis on likely scenarios for small odontocetes taken as by-catch in fisheries.

4.1.2 Abundance

Performance will depend on absolute abundance, and on the split of abundance between stocks. All archetypes should have the same total pristine abundance and it was **agreed** that initial scenarios should use 7,500, 15,000 and 30,000 animals. A total abundance of 30,000 is likely to prove taxing for genetic simulation models, but it is important to use numbers that will be relevant to likely applications. The abundances of the different stocks should be kept equal for Archetypes IV and V. Both equal and unequal abundance scenarios will be examined for Archetype II. In the unequal abundance scenario where there are two populations, 90% of the abundance will be in one population and 10% in the other. In the unequal abundance scenario with three populations, 45% of the abundance will be in one edge population, 45% in the middle population and 10% in the other edge population.

4.1.3 Dispersal

When abundances are unequal, there is a question over how to model dispersal in such a way that the differences in abundance are maintained. One possibility is simply to choose different rates depending on the area. However, it was argued that this approach would be inappropriate because it implies different biology (i.e., a different innate dispersal tendency) in different populations. It would also fail to produce the genetic effect that the unequal abundance scenario is designed to capture: the higher rate of genetic drift in a small population being swamped out by gene flow from a large neighboring population. In genetic models, this issue is typically dealt with by using the same dispersal rate in all populations and using a "hard" carrying capacity (excess animals disappear) to maintain the appropriate abundances in all populations. It was **agreed** that this approach should be used in the initial set of TOSSM simulations. More complicated density-dependent models of dispersal may be considered at later stages.

The first phase of simulations would use four dispersal rates, ranging from one migrant per generation to 0.5% per year, and the others equi-spaced on a log scale. For Archetype III, dispersal rates will need to be adjusted to achieve comparable rates over the whole population range to the rates in Archetype II; keeping the dispersal rate constant and increasing the number of substocks would miss the point.

4.1.4 Mating system

Polygyny can have a large impact on effective population size. For the first phase of simulation, the Workshop **agreed** to consider only lottery polygyny, as this is straightforward to simulate. Future simulations might need to include more complexity, including individual male histories in order to allow a realistically-skewed distribution of male reproductive success.

4.1.5 Biological parameters

In order to get the most insight into the comparative behaviour of boundary-setting methods, it was **agreed** that for to begin with an MSYR of 4% should be used, and initial depletions (where appropriate) of 30%.

4.1.6 For the future

The Workshop recognised that this would be an iterative and long-term project. It briefly considered what (more complex) archetypes might be considered in the future. Some of those raise included: two-dimensional patterns of stock separation; partial separation of sexes; inter-annual shifts in feeding ground location; pulsed rather than constant dispersal rates; density-dependent dispersal; seasonal sympatry of a resident and a migratory population. The question of choosing the most appropriate future archetypes should wait until the results from the 'first phase' are available.

4.2 Simulating the collection of genetic data

Although some methods can work with gene-frequency data aggregated across individuals, other methods require that each sample be individually recorded and associated with a particular sampling location. Consequently, it was **agreed** that the simulation of genetic samples should be spatially explicit; i.e. each sample should be associated with a specific space/time location.

It was **agreed** that the default sampling scheme would be systematic random sampling (the ideal towards which all studies should strive). However, it was recognised that real datasets often include several types of sampling bias (e.g. sampling of relatives, spatially clustered sampling, sampling gaps etc.) and that these should be investigated. It was **agreed** that sampling gaps should be the prime variant to examine in the first stage of the simulations. Spatio-temporal biases should be considered at a later stage.

There was some discussion of whether samples should be taken from breeding or feeding grounds. It was **agreed** that although eventually sampling on both feeding and breeding grounds would need to be considered, initial efforts should focus on sampling from the feeding grounds, which will usually present a greater challenge to the analytical methods. The possibility that sampling on the breeding grounds will improve performance when harvesting is only on the feeding grounds warrants further investigation.

It was **agreed** that in the first phase, all samples should be collected in the same year (i.e. ignoring temporal variation). It was noted that this would preclude the use of standard genetic tagging methods. There was some discussion of how tagging methods could be applied to a sample from a single year by using information from first and second order relatives detected in the sample. This warrants further consideration.

There was some discussion of appropriate sample sizes to be simulated and whether they should be absolute or proportional to the population size. It was suggested large (500) sample sizes should be included, since there are some very large sample sets available. Others believed that large samples are not important, particularly for analytical methods based on coalescent theory, and that time and energy would be better spent increasing the number of loci. It was noted that proportional sampling is more important for some methods (e.g., frequency-based methods) than others. Finally, it was agreed that in the first phase, absolute sample sizes of 50 and 100 per population should be considered (I assume this number is per population). Proportional sampling and larger sample sizes should be examined in the future. It was recognised that eventually it will be important to investigate the effect of changing sample size and the number of loci, on the performance of different methods.

4.3 Spatial harvest strategy

The workshop **agreed** to use the CLA to assign quotas according to two different RMP variants (see below). This does not restrict applicability to commercial whaling situations; in principle, any reasonable catch-control rule could be used, but the CLA is well-understood and fast programs are already available. In order to provide maximum insight into the likely management performance of boundary-setting methods, the workshop **agreed** to consider catch strategies that would place catches as close as possible to whatever management boundaries were chosen, rather than evenly spread out between boundaries. Punt commented that a spatially-explicit continuously-distributed version of the population dynamic equations already used in the IWC could be developed, and offered to provide a paper to the Scientific Committee on this issue. The Workshop **welcomed** Punt's offer.

It was agreed to trial two different RMP variants. In the first, management boundaries are treated as defining *Medium Areas,* so that quotas are set independently for each area depending only on that area's abundance estimate. In this version, total catch will be greatly reduced if too many boundaries are selected, because of the precautionary response of the CLA to increased uncertainty about abundance. In the second variant, catch-cascading will be applied by basing total quota on the abundance estimate for the whole region (which will be relatively more precise than for individual areas), and then the total quota is pro-rated across the selected management areas. Performance statistics will include some measure of the logistical implications for whaling of different boundary selections (section 6.5).

5 DETAILS OF GENETIC SIMULATION

5.1 Equilibrium and non-equilibrium models

The theory and simulation of population genetics is simpler when model populations are assumed to be in mutation-migration-drift equilibrium. However, whale populations (even pre-exploitation) may not be in equilibrium, and an assumption of equilibrium in the simulations may give unrealistic results. For one thing, if simulations are restricted to populations that are in equilibrium, then an unfair advantage will be given to methods that assume equilibrium *a priori*. The problem with non-equilibrium simulations, however, is that it is necessary to choose a particular non-equilibrium state out of many possibilities. The Workshop **agreed** that it would eventually be important to investigate non-equilibrium populations, but that this should be deferred until Phase 2.

5.2 Models for genetic simulations

Three commonly-used types of models were discussed: Individual-Based Models (IBMs), birth-death models, and coalescent models. Coalescent models have the advantage of being much faster than the other two, but also have some necessarily rigid assumptions; for example, they cannot be adapted to allow for differential mating success. It was noted that a coalescent model would be unable to handle Archetype V. Birth-death models much more flexible than coalescent models and incorporate fewer assumptions, but they are still limited in terms of the degree of biological realism that can be achieved. IBMs are the slowest models, but have the crucial advantage that they can easily be modified to introduce whatever level of realistic complexity is desired (e.g. other mating structures). Even if such complexities are excluded from the initial round of simulation trials, it is important not to set up a structure that will seriously constrain future investigations. IBMs are also the only way to simulate data for kinship-based methods. An attractive possibility is to run coalescent simulations up to, say, 1000 generations before the present, to generate "recent" allele frequency distributions, and then to allow IBMs to bring the populations up to the present, e.g. under more realistic mating structures. This would blend the realism of an IBM with some of the speed advantages of a coalescent. However, it is important to check that this is an acceptable approach by investigating whether coalescent

simulations do generate similar numbers of haplotypes, similar Fst and Phist values, and similar frequency distributions of repeat lengths across alleles, to long-term IBMs. It will also be important to make sure that the simulated values of these variables are broadly in line with what is seen in whale populations.

The Workshop **agreed** that the first step should be to compare the results of IBMs initialized with (1) random haplotypes; (2) observed haplotype/allele frequency distributions from natural populations; and (3) data generated by a coalescent. If allowed to run long enough, simulations using all three initialization methods would eventually come to equilibrium and produce the same results. The practical question is how long it takes in real time for the model to reach equilibrium with each initialization scheme. Beerli offered to make available his coalescent-generating code for this exercise. There are also existing IBM programs that could be used for at least some aspects of the simulations (such as Easypop, Metapop and Popsim), although substantial modifications may be necessary.

5.3 Genetic variables to simulate

When using the coalescent to generate initial allele/haplotype distributions, simulating large population sizes is possible but slow. It was suggested that the number of samples a coalescent model would need to simulate in order to represent the population, is of the order of the square root of the effective population size. It was also suggested that for simulating a population with a census size of 30,000 using a coalescent, it would be sufficient to simulate only 1,000; the gene frequencies of the 1,000 sampled individuals should adequately represent the those of the entire population.

The Workshop **agreed** that simulations should cover four types of genetic markers: mtDNA, microsatellites, multilocus sequencing and SNPs. MtDNA and microsatellites are currently used on a routine basis in studies of population structure, while multilocus sequences and SNPs are relatively new technologies that are likely to become increasingly important in the future. The number of microsatellite loci to simulate was discussed. Current whale studies use only 8-12 loci, but it was noted that there are some questions that can only be addressed with much larger numbers, and that logistically it is getting easier to analyze large numbers of loci. The Workshop **agreed** that it would be useful to simulate both small (10) and large (30, 100) numbers of loci. Once a large number of loci have been simulated, of course, a smaller subset can be used in analysis. Both allele frequency and allele size should be recorded.

Simulations of mtDNA should include approximately 40 variable sites, as this is a typical number found in most species. It was suggested that simulations for multilocus sequences could include 10-300 sequences and that SNPs could be retrieved from those sequences.

The participants agreed that it was difficult to obtain independent estimates for the effective population size (Ne) and the mutation rate (μ). However, at least for populations at equilibrium, allele distributions are actually determined by the product theta=4*N_e*μ, rather than N_e and μ individually. Plausible values for theta can be established by comparisons with real data.

Realistic rates of laboratory error (both sequencing error, and errors in the scoring of microsatellite alleles) should be incorporated into the simulated datasets. Norwegian scientists have sent identical DNA samples to several independent labs, and overall about 1% of reported alleles differ between the labs. This type of error is allowed for in some existing genetic simulation programs.

6 CHOICE OF PERFORMANCE MEASURES

6.1 Testing methods that require *a priori* inputs

Many of the methods in Table 1 need to be given a set of possible boundaries, before deciding how many of those boundaries ultimately need to stay. Conversely, some of the methods need to be "told" how many boundaries should be placed, and will then choose where to put those boundaries. It is both important and difficult to fairly test those methods which only provide a partial solution to the problem of defining management. The group **agreed** that methods could be fairly tested by breaking the performance testing into three distinct tests:

(1) Given that the number of stocks is known, does a method accurately stratify the data? This test would only apply to methods that are designed to stratify a dataset. Each analyst would be told the actual number of stocks represented by the sample and would use their assigned method to find the best way to stratify the data into that number of strata.

(2) Given a pre-stratified dataset, does the method define the correct number of stocks? In this test, a method would be provided with a dataset that is already divided into, say, three hypothesized stocks. The method would then need to determine whether the area should be managed as one, two or three stocks.

(3) Given an un-stratified dataset and no information on the actual number of stocks represented by the dataset, how well does the method define stocks? This is the most complete, and therefore the most challenging, test. At present, most of the analytical methods are not capable of both stratifying the data and choosing the appropriate number of stocks.

Tests 1 and 2 will reveal how well each method performs the function it is designed to perform (either stratify the data or evaluate *a priori* hypotheses regarding population structure). However, because both tests provide information that would not be available in a real-world setting, the results will be of limited value in determining how well a method would perform in reality.

Thus, it is test 3 that is of greatest interest. Two methods of implementing Test 3 for methods that are not capable of both stratifying the data and choosing the appropriate number of stocks were discussed. For methods requiring pre-stratification, one possibility is to ask an external expert/s who is well-versed in cetacean biology, behavior and population structure but naïve to the archetypes represented in the simulated datasets to perform the *a priori* stratification. The expert/s would be given some basic ancillary information of the type that is generally available to researchers when they decide how to stratify their samples (e.g., a map of sample locations, rough distributional information, bathymetry, general oceanographic information, etc.) to assist them in deciding how to stratify the data. It was **agreed** that further discussion of this possibility would be necessary in order to determine the feasibility of this approach, decide what types of ancillary data the expert/s would be supplied with and to generate a list of people who could be asked to fill this role.

A second approach might be to combine a method that is designed to stratify datasets with one that is designed to evaluate pre-stratified datasets. For example, the hypothesized stocks generated by Boundary Rank could be evaluated using Migrate. Many such combinations of methods are possible. Combining methods in this way could result in bias, since the same data would be used twice. However, this might not matter for management, since the implications of any such bias would be revealed in the simulation tests.

6.2 Absolute performance measures

The primary goal of the project is to evaluate the performance of different boundary placement methods in the context of the regulation of whaling. Consequently, most of the performance measures used in the exercise will focus on how well the stocks defined by a particular method meet the management objectives set out by the Commission. In addition to these whaling-related performance measures, however, it will also be helpful to develop and record "absolute" performance measures that are specific to each analytical method. Most existing methods of analyzing population structure were not designed specifically for setting management boundaries. The absolute performance measures would evaluate how well each method served the purpose it was originally designed to serve. The results of these absolute performance measures will be of great value, since most methods have not been subjected to extensive performance testing in any context. Since most methods will need some further development before they can be used in a management context, feedback on the performance of the core method may be helpful when extending the functionality of the method to management issues. The appropriate choice of absolute measures would depend on the original purpose each method was designed to serve, and is therefore left to individual developers.

6.3 Management-based measures

The RMP is designed around qualitative objectives related to conservation (level of depletion) and utilisation (levels of catches, stability of catches). To assess how well these objectives are likely to be met by particular catch-control rules, the Scientific Committee devised a number of quantitative performance measures that summarize the results from simulation trials. Each performance measure is relevant to a different aspect of the objectives. For example, one performance measure aimed at evaluating over-depletion is the 5th lowest level of depletion encounter in 100 simulations. Similar objectives have been set for aboriginal whaling management, and appropriate performance measures developed.

These same qualitative objectives still apply when testing boundary-setting methods, so TOSSM can use existing performance statistics for total catch. It should not be necessary to consider stability of catches, because the catch-control rules which would be applied after boundary-setting are already designed to achieve this. There are several aspects of performance that do require further consideration, as discussed below.

6.3.1 Unit-to-conserve; assessing local depletion

Population structure can occur on many different scales. At one end of the spectrum there is near-complete isolation between sub-species, while at the other end there may be small social units between which gene flow is very high. Local depletion of a family group might be quite acceptable for management; "local" depletion spanning half an ocean basin might not be. The spatial scale is closely linked to the time scale for potential recovery; a local feeding ground along a coastline might be recolonized within a decade, whereas refilling half an ocean basin might take centuries. There are also some complex issues concerning possible local depletions in parts of the annual feed-move-breed cycle that cannot be directly observed; is it only important to conserve phenomena that can be directly observed?

Deciding which scale to manage at– in other words, defining the unit to conserve – is a job for managers and policy-makers, not scientists. However, there is as yet no clear guidance for scientists on what managers (i.e. the IWC) want in

this respect; different managers may have different opinions, and managers might modify their opinions after hearing scientific results. The scientists' job is therefore to report results on a variety of scales, leaving the ultimate choice to the managers.

The simplest and most easily interpreted way to record depletion on different scales, is to start by dividing up the area being simulated into numerous cells (see section 4.3), and to keep track of depletion within each cell. This applies to unharvested breeding grounds as well as to harvested feeding grounds. Individual animals may (in the simulation, and in reality) have a complex mix of movement patterns between breeding and feeding grounds, but as long as all cells remain well-filled, then there is obviously no serious local depletion. Cells can be aggregated to provide reports at different spatial scales. Numerical measures of depletion on a per-cell (or cell-group) basis can be adapted from statistics already used elsewhere in the IWC.

6.4 Genetic measures of depletion

If a subpopulation becomes severely depleted, its long-term viability might be impacted more than the abundance drop alone suggest, because ofa genetic loss that compromised the population's ability to adapt to future biotic or abiotic challenges. Fortunately, for most populations to which these studies are potentially relevant, the inherent conservatism of the catch-control rules is such that there is unlikely to be depletion down to genetic bottleneck population sizes (a few hundred animals). It is not possible to identify a specific threshold of loss that corresponds to compromising evolutionary potential. However, assuming the principle that more genetic diversity is generally "good" the TOSSM exercise can produce measures of genetic "health" that allow comparisons among the different methods. Because the statistic for genetic diversity (H) is a fairly crude measure, the group agreed that the starting and ending number of alleles after a 100 year management program and the proportion of rare alleles (present in <1% of individuals) would be useful measures.

It was noted that PVA ("population viability analysis") simulations of small populations, have sometimes shown major viability failures only after 200 or 300 years; it might be worth extending the time period of some of the simulations beyond the traditional 100 years.

6.5 Measuring the logistic consequences of boundary placement

A plausible argument can be made that good conservation and good catches will both result from the following simple approach to boundary placement and quota setting: calculate total quota ignoring any boundaries, divide up the range very finely into a large number of areas, and pro-rate the total catch in proportion to each area's abundance ("extreme catch cascading"). The downside of such a policy is that it places a major logistic constraint on whalers, which in some cases might prevent the fishery from filling its quota or even operating at all. When evaluating the performance of different methods of defining stocks, there needs to be some way of penalizing methods that set more boundaries than are needed for effective conservation (of whatever unit-to-conserve is chosen). There are inevitable trade-offs between logistics and conservation here; how to make those trade-offs is again an issue for managers, not scientists. What scientists have to do is report how a given analytical method has affected the distribution of whaling effort in space and time. This needs to be summarized into a performance measure of some kind. For the semi-abstract simulations that will be undertaken initially, it is easy to think up various simple candidates (mean distance from "land"; proportion of effort within X km of "land"; etc.). The task of settling on an appropriate performance measure for an initial round of simulations, was deferred to the Stock Definition subcommittee.

7 CONCLUSIONS AND THE WAY FORWARD

During the workshop, considerable progress was made in developing the TOSSM project. Nonetheless, it was recognized that many details of the simulations and testing procedure still remain to be specified. In order to structure the work, six distinct 'modules' were identified: (i) genetic population simulation, (ii) specification of population dynamics (archetypes), (iii) specification of genetic sampling, (iv) specification of catch strategy, (v) adapting methods for automated use in management context, and (vi) integration of these components into an initial set of trials. Work on (i)-(v) can proceed largely in parallel; the amount of work that will be required varies greatly between the modules, e.g. with number (ii) being largely completed during the workshop. The modules are discussed further below, followed by a proposed timeline. It was **agreed** that the specification of remaining details should start during the Scientific Committee in Berlin, where a TOSSM Steering Group (distinct from the workshop steering group) should be established to co-ordinate work intersessionally.

The group **noted** that the six modules below would constitute only the first phase of TOSSM. This phase is aimed at elucidating the basic comparative and absolute properties of boundary-setting methods, and at encouraging method

development. It is envisaged that there will be a second phase, in which (ii)-(iv) are reconsidered in more detail (e.g. more archetypes, adaptive management with changing boundary placement, etc.).

7.1 Modules for further work

7.1.1 Genetic simulation

The most fundamental and difficult task is the coding and validation of a program for simulating the genetic datasets to be used in performance testing. There are several genetic individual-based models (IBMs) available, one of which may be able to serve as the basis for the TOSSM program. If that is the case, the development time and cost can be considerably reduced. It was recognized that none of the existing programs is likely to incorporate all of the features required, so some augmentation will be needed.

Irrespective of whether the TOSSM program is completely new or based on an existing program, the participants **agreed** that the performance of the program should be thoroughly examined to ensure that it is working properly. This validation should include comparison of the model output to (i) theoretical expectations, (ii) other genetic models, and (iii) empirical datasets. The last will allow the model's mutation parameters to be tuned so that the simulated data are consistent with the genetic characteristics actually observed for marine mammal species.

Participants identified three means by which the haplotype and allele frequencies within an IBM could be initialized: (i) generation by a coalescent model, (ii) choosing from observed allele frequency distributions, or (iii) random assignment to individuals from an infinite allele model (see section 5.2). It was **agreed** that the implications of these three initialization methods should be examined to determine whether they all produce the same results and what type of speed gains can be expected from using one of the first two methods of initialization.

The workshop **recommends** that Luikart's research group should take the lead in developing the model, due to their experience in designing and evaluating genetic IBMs. David Tallmon, a post-doctoral fellow in the Luikart laboratory, was identified as a likely candidate for acting as lead modeller. Ralph Tiedemann and Karen Martien volunteered to assist in the modelling effort. A small group was formed to work intersessionally on investigating existing IBMs for their appropriateness in the TOSSM exercise. The group members are Gordon Luikart, David Tallmon, Barbara Taylor, Karen Martien (convenor) and Ralph Tiedemann.

7.1.2 Specifying biology and population dynamics

The workshop **developed** a list of general archetypes for initial consideration (section 4.1.1), together with some more detailed proposals for population dynamics, reproductive biology, and biological parameters (section 4.1). The workshop **agreed**, at least for Phase I, that population dynamic models should correspond to those already in use within IWC.

7.1.3 Specifying the genetic sampling scheme

It was **agreed** that, during phase I of the project, two different genetic sampling schemes should be simulated: random sampling, and a scheme in which there were gaps in the spatial distribution of the samples. Random sampling is the ideal toward which all studies should strive, but gappy sampling is common and is a possible cause of bias. It was **agreed** that the task of specifying the details of these sampling schemes and determining how they should be implemented within the model would be assigned to a small working group to be formed at the SC meeting in Berlin.

Initially, only simulated feeding-ground samples will be made available. In phase 2, it is planned to incorporate breeding ground samples too, plus such biasses as a tendency to sample related animals. Proposals concerning number of loci, number of samples, etc., are given in section 4.2.

7.1.4 Catch strategy

The workshop **agreed** to use the CLA to assign quotas according to two different RMP variants (see below). This does not restrict applicability to commercial whaling situations; in principle, any reasonable catch-control rule could be used, but the CLA is well-understood and fast programs are already available. In order to provide maximum insight into the likely management performance of boundary-setting methods, the workshop **agreed** to consider catch strategies that would place catches as close as possible to whatever management boundaries were chosen, rather than evenly spread out between boundaries. Punt commented that a spatially-explicit continuously-distributed version of the population

dynamic equations already used in the IWC could be developed, and offered to provide a paper to the Scientific Committee on this issue. The Workshop **welcomed** Punt's offer.

It was agreed to trial two different RMP variants. In the first, management boundaries are treated as defining *Medium Areas,* so that quotas are set independently for each area depending only on that area's abundance estimate. In this version, total catch will be greatly reduced if too many boundaries are selected, because of the precautionary response of the CLA to increased uncertainty about abundance. In the second variant, catch-cascading will be applied by basing total quota on the abundance estimate for the whole region (which will be relatively more precise than for individual areas), and then the total quota is pro-rated across the selected management areas. Performance statistics will include some measure of the logistical implications for whaling of different boundary selections (section 6.5).

7.1.5 Adapting methods for automated use in a management context

The workshop identified three components of a boundary-setting method that are necessary for automated testing:

1. Generation of a set of possible boundaries ("the front end").
2. Evaluating the possible boundary combinations.
3. Interpreting the results from step (2) to end up with a set of boundaries for management ("the back end").

As discussed in section 3.7, most of the available methods only perform either step 1 or step 2. The group **agreed** that attention would need to be given to the development of some generic "front ends" and "back ends", although method developers would of course be free to devise their own. In practice, decisions about *a priori* stratification are often made on the basis of additional non-genetic data, but this might be diffficult to reproduce in a simulation-testing framework. It was **concluded** that further attention should be given to this issue during IWC Berlin.

'Back-ends' will generally consist of a 'threshold' for deciding whether the value of the statistic output in step 2 is high or low enough (as appropriate) to warrant placing a boundary. This threshold could, for instance, be a dispersal rate below which two areas should be managed separately. The threshold value is an adjustable parameter of each method, and the process of simulation testing will enable developers to choose threshold values that achieve effective management results. For methods geared towards single hypothesis tests, the development of an automated "back end" that can sift through vast numbers of p-values is likely to be quite challenging.

7.1.6 Integrating the modules to form an initial set of trials

The workshop **recognized** that, once 7.1.2-7.1.5 have been dealt with, it will be necessary to decide which cross-combinations of biology, sampling, and management options should be trialled in Phase I of TOSSM. One lesson from other simulation testing exercises is the importance of avoiding an overwhelmingly large number of trials until and unless it is absolutely necessary.

The precise format of simulated datasets will be need to be specified; this is the one pre-workshop objective that was not specifically covered by the workshop, but should not be a major difficulty. The workshop **recommended** that the simulated data be made publicly available, preferably over the Internet. Further, an "RMP-lite" program should also be made available, that will apply the CLA across the chosen boundaries, collect performance statistics (section 6), and allow developers to test management performance themselves. In this first phase of TOSSM, the workshop **agreed** that it would be most useful to reveal what simulation details were used to generate each dataset, as this is likely to help in developing better methods. When developers have gained more experience, it will be desirable to include blind-trial datasets where the truth is not made available to developers; this is the procedure followed, for example, in the Abundance Estimator Testing of the IA Standing Committee.

The workshop **noted** that the datasets generated would be of considerable interest to genetic population modelling in general, aside from their particular value for whaling management simulations.

7.2 Proposed timeline for Phase I

Developing the genetic simulator will be by far the most time-consuming of the above tasks; 4-6 months' work might be required. Since this highly specialized work will have to be done by someone currently outside the IWC, and the start date is unknown, it is not possible to be sure exactly when this module will be finished. However, work on all the other

aspects 7.1.2-7.1.5 can be completed in the meantime by regular IWC participants. It should be possible to address 7.1.6, and to have simulated datasets ready, in time for IWC 56 in 2004.

Issues to be addressed in Phase II, and the time required to address them, will be determined iteratively after seeing results for Phase I, as is normal in management procedure evaluations.

7.3 Encouragement of further methods development

During the workshop, mention was made of several other analytical methods whose developers were not at the Workshop. The Workshop **emphasized** the importance of contacting these developers and inviting them to participate in the performance testing exercise. The further refinement of existing methods was strongly **encouraged**, preferably in collaboration with IWC scientists, and the development of new methods of investigating population structure within a management context. It was further **agreed** that the dissemination of information and co-ordination of work should be handled by the Intersessional Steering Group to be established during the IWC Scientific Committee in Berlin. An important function of the TOSSM project, besides evaluating *existing* methods in the context of whaling management, is to draw attention to the management and conservation implications of population structure in general, and to inspire more development of practical methods.

LITERATURE CITED

Archer, F. 2000. Centroid: Analysis of Population Overlap Using Principal Coordinates on Genetic Distance Data. In Preparation.

Beerli, P. and J. Felsenstein. 1999. Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. Genetics. 152: 763-773.

Beerli, P. and J. Felsenstein. 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. Proceedings of the National Academy of Science. 98(8): 4563-4568.

Berthoud F., P. Taberlet and S. Manel. In review. Genbar : a program to identify genetic boundaries from multilocus genotype. Submitted to The Journal of Heredity.

Cassens, I, R. Tiedemann, F. Suchentrunk and G. B. Hartl. 2000. Mitochondrial DNA variation in the European Otter (Lutra lutra) and the use of spatial autocorrelation analysis in conservation. Journal of Heredity. 91(1):31-35.

Cui, G., Punt, A. E., Pastene, L. A., Goto, M. 2002. Bayes and empirical Bayes approaches to addressing stock structure questions using mtDNA data, with an illustrative application to the north Pacific minke whales. *Journal of Cetacean Research and Management* 4:123-134.

Dupanloup, I., Schneider, S., and Excoffier, L. 2002. A simulated annealing approach to define the genetic structure of populations. *Molecular Ecology* 11:2571-2581.

Hudson, R.

Martien, K.K. and Taylor, B.L. In review. A new method using genetic data to generate hypothesized population structures for continuously distributed species. Submitted to *Journal of Cetacean Research and Management.*

Monmonier, M.S. 1973. Maximum-differences barriers: an alternative numerical regionalization method. *Geographical Analysis* 3:245-261.

Moritz, C. 1994b. Defining 'evolutionarily significant units' for conservation. *Trends Ecol. Evol.* 10:373-5.

Pritchard, J.K., Stephens, M. and Donnelly, P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155:945-59.

Roff, D.A., Bentzen,P. 1989. The statistical analysis of mitochondrial DNA polymorphisms: χ^2 and the problem of small samples. *Molecular Biology and Evolution* 6:539-45.

Ryder, O. 1986. Species conservation and systematics: the dilemma of subspecies. *Trends Ecol. Evol.* 1:9-10.

Strand, A. E. 2002 METASIM 1.0:an individual-based environment for simulating population genetics of complex population dynamics. Molecular Ecology Notes 2:373-376.

Templeton, A. R., E. Routman and C. A. Phillips. 1995. Separating population structure from population history: a cladistic analysis of the geographic distribution of mitochondrial DNA haplotypes in the tiger salamander, Ambystoma tigrinum. Genetics. 140: 767-782.

Tiedemann, R. O. 2001. Stock identification in continuously distributed species using molecular markers and spatial autocorrelation analysis. Paper SC/53/SD3 submitted to the Scientific Committee of the International Whaling Commission.

Womble, W.H. 1951. Differential Systematics. *Science* 114:315-322.

]

TABLE 1.

REQUIREMENTS FOR DIFFERENT APPROACHES TO IDENTIFYING POPULATION STRUCTURE

METHOD	INPUTS	ASSUMPTIONS	ORIG. AIM	EXTRAS NEEDED	SIM. MODEL
(generic)	(nuclear DNA needed by most)	no linkage no nulls representative samples no genotyping errors random mating last N gens representative of now			(and what if we want to simulate MORE genetic samples during the 100 years of simulation?)
Hypothesis testing	hap. freqs by stratum a priori strata		test separateness	threshold p-val pre-stratifier	group birth/death
"Bayesian hypo testing"	as above	allele priors	test separateness	threshold BFactor pre-stratifier	group birth/death
STRUCTURE etc.	indiv. genotype	prior on K	indiv. assignment	link to geography of samples	group birth/death
MIGRATE etc.	indiv. genotype a priori strat mutation model	Evolutionary history	migration rates	threshold mig rate pre-stratifier	IBM
B RANK	hap. freqs by site connectivity		hiearchical clusters	threshold in hierarchy	group birth/death
Spatial autocorr	indiv. genotype		is there IBD? are there steps	thresholding...	IBM
CENTROID	indiv. genotype		indiv. assignment + test separateness	pre-stratifier thresholder	IBM
N clade				too much is "by hand" for now	
Genetic tagging	indiv genotypes	no family structure?	estimate current dispersal	estimation method!	IBM

Annex A

List of Workshop Participants

Scott Baker – University of Auckland. Auckland, New Zealand.
Peter Beerli – University of Washington. Seattle, WA USA.
Peter Best – South African Museum. Cape Town, South Africa.
Mark Bravington* – CSIRO. Hobart, Australia.
Doug Butterworth – University of Cape Town. Cape Town, South Africa.
Anna Danielsdottir – Marine Research Institute. Reykjavik, Iceland.
Greg Donovan* – Head of the IWC Secretariat. London, England.
Phillip England – Université Joseph Fourier. Grenoble, France
Rus Hoelzel – University of Durham. Durham, UK.
Naohisa Kanda* – Institute for Cetacean Research. Tokyo, Japan.
Toshihide Kitakado – Tokyo University of Fisheries.
Gordon Luikart – Université Joseph Fourier. Grenoble, France
Karen Martien* – Southwest Fisheries Science Center. La Jolla, CA, USA
Scott Mills – University of Montana. Missoula, MO, USA.
Per Palsbol – University of California, Berkeley. Berkeley, CA, USA.
Luis Pastene – Institute for Cetacean Research. Tokyo, Japan.
Jonathan Pritchard – University of Chicago. Chicago, IL, USA
William Perrin – Southwest Fisheries Science Center. La Jolla, CA, USA
Andre Punt – University of Washington. Seattle, WA, USA
Howard Rosenbaum – I need to get Howard's affiliation off the website
Hans Julius Skaug* – Institute of Marine Research. Bergen, Norway.
Barb Taylor* – Southwest Fisheries Science Center. La Jolla, CA, USA
Ralph Tiedemann – Universitaet Potsdam. Berlin, Germany.
Lars Walloe – University of Oslo. Oslo, Norway.

* - Steering committee member

Annex B

Workshop Agenda

1. Opening remarks
2. Election of Chair and appointment of Rapporteurs
3. Background presentations
   - 3.1 Introduction to simulation performance testing
   - 3.2 Review of whale biology
     - 3.2.1 General demography and life history
     - 3.2.2 Case studies: migratory patterns, population structures and exploitation
     - 3.2.3 Typical examples of available genetic data
   - 3.3 Whaling management procedures and the role of simulation performance testing in an applied setting
   - 3.4 Introduction to performance measures and possible definitions of the "unit to conserve"
   - 3.5 Introduction to genetics
     - 3.5.1 Marker types
     - 3.5.2 Ways to model genetic systems (Birth-death, coalescence or individual-based-models; Stepping-stone or diffusion)
   - 3.6 Review of analytical methods for investigating population structure[2]
4. Choose scenarios to simulate
   - 4.1 Select archetypes
   - 4.2 Choose sampling schemes
   - 4.3 Choose spatial harvest strategy
5. Develop simulation models
   - 5.1 What type of models to use? (birth-death, coalescence, IBMs)
   - 5.2 Choose genetic parameters (marker types, number of markers, mutation rate and model, mating system)
6. Develop performance measures
   - 6.1 Absolute measures
   - 6.2 Management-based measures
     - 6.2.1 Performance under the RMP
     - 6.2.2 Performance under the AWMP
   - 6.3 How do we fairly test methods that require *a priori* stratification?
7. Set-up project structure and timeline
   - 7.1 Specify content and format of simulated datasets
   - 7.2 Should the simulation modelers be separate from the population structure testers?
   - 7.3 Project timeline
     - 7.3.1 Should the exercise be divided into phases (e.g., mtDNA first, nuclear DNA second)?
     - 7.3.2 What can be accomplished prior to May?
     - 7.3.3 Establish long-term timeline

[2] AMOVA-type hypothesis testing, Structure (Pritchard), Migrate (Beerli), Boundary Rank (Martien), spatial auto-correlation (Tiedemann), Bayes Factors (Punt), nested clade analysis, genetic tagging

Annex C

List of Documents

Beerli, P. and J. Felsenstein. 1999. Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics*. **152**: 763-773.

Beerli, P. and J. Felsenstein. 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in *n* subpopulations by using a coalescent approach. *Proceedings of the National Academy of Science*. **98**(8): 4563-4568.

Cassens, I, R. Tiedemann, F. Suchentrunk and G. B. Hartl. 2000. Mitochondrial DNA variation in the European Otter (*Lutra lutra*) and the use of spatial autocorrelation analysis in conservation. *Journal of Heredity*. **91**(1): 31-35.

Cooke, J. G. 1994. The management of whaling. *Aquatic Mammals*. **20**(3): 129-135.

Cui, G., A. E. Punt, L. A. Pastene and M. Goto. Bayes and empirical Bayes approaches to addressing stock structure questions using mtDNA data, with an illustrative application to North Pacific minke whales. *Journal of Cetacean Research and Management*. **4**(2): 123-134.

Donovan, G.P. 1991. A review of IWC stock boundaries. *Report to the International Whaling Commision* Special Issue 13:39-68.

Martien, K. K. The impact of the unit to conserve on designing performance measures for TOSSM.

Martien, K. K. and B. L. Taylor. A New Method for Using Genetic Data to Define Management Stocks for Marine Mammals. In review at *Journal of Cetacean Research and management.*

Martien, K. K. and B. L. Taylor. 2002. Discussion paper to design performance trials for the Testing of Spatial Structure Models. Paper SC/54/SD6 submitted to the International Whaling Commission.

Palsboll, P. J. 2001 The JARPN 2000 workshop crash-course in cetacean genetics. *Journal of Cetacean Research and Management*. **3**(suppl.): 399-407.

Pritchard, J.K., Stephens, M. and Donnelly, P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155: 945-59.

Taylor, B. L. Testing performance through simulations.

Taylor, B. L. Population structure archetypes.

Taylor, B. L. Connecting management, population dynamics and genetics for TOSSM.

Taylor, B. L. and A. E. Dizon. First policy then science: why a management unit based solely on genetic criteria cannot work. *Molecular Ecology*. **8**: S11-S16.

Tiedemann, R. O. 2001. Stock identification in continuously distributed species using molecular markers and spatial autocorrelation analysis. Paper SC/53/SD3 submitted to the International Whaling Commission.

Tiedemann, R., O. Hardy, X Vekemans and M. C. Milinkovitch. 2000. Higher impact of female than male migration on population structure in large mammals. *Molecular Ecology*. **9**: 1159-1163.

Annex D

Overview of large whale biology

(Peter Best and Bill Perrin)

For the benefit of those participants unfamiliar with marine mammal biology, a brief overview of general demography and life history was presented, concentrating on baleen whales and the sperm whale as being the species with which the IWC had historically been most concerned.

Most baleen whales migrate seasonally over long distances between winter calving and breeding grounds in subtropical/tropical waters and polar feeding grounds in summer. For the best-known species, the humpback whale, the breeding grounds tend to be concentrations of animals in shallow waters near coasts or island groups, and animals from one breeding ground can migrate to different feeding grounds. Fidelity to a particular feeding ground is maintained through the calf accompanying its mother during a lactation period of 6 months to a year. Individuals disperse widely when on the feeding ground, with varying degrees of intermingling between animals from different breeding grounds (= mixing). Permanent transfer of an individual from one breeding ground to another (= dispersal) can occur. The information for other baleen whales (apart from the gray whale) is much less, and the location and nature (concentrated, dispersed) of the breeding grounds are frequently unknown. Baleen whale migrations are structured by sex, age and reproductive status, such that arrival and residence times on breeding and feeding grounds can differ substantially depending on these factors.

Unlike baleen whales, sperm whales display a marked social organization, with females and their offspring forming separate groups from most subadult and adult males. Although males migrate in summer into polar waters, female-led groups remain in warmer waters year-round. Feeding and breeding "grounds" therefore do not exist as separate entities, and the species occurs pelagically throughout each ocean basin. Nevertheless, the ranges covered by individual females are relatively limited (usually <500 n. miles), and much less than those of males (up to 2,000 or 3,000 n. miles). Although the basic social unit of females is small (10-12 animals) and believed to persist for years to decades, more temporary aggregations of up to several hundred animals can occur in response to local environmental conditions, and over ranges of 100s to 1,000 km. Differences in apparent conception dates in southern sperm whales suggest a "cline" in the peak breeding season from the SW Atlantic to the SE Pacific. Significant differences occur at a scale of 40 degrees of longitude, indicating that some form of population structure may exist, even within one ocean basin.

Large whales in general are long-lived, with average life expectancies ranging from 16 to over 100 years, and have low reproductive rates owing to small litter sizes and long calf dependency. Maximum population growth rates are therefore low (3-10%), and have historically easily been surpassed by annual fishing rates, resulting in declines to a few percent of original abundance in several cases.

The IWC presently recognizes 15 species of great whales, 14 baleen whales and one toothed whale, as follows:

Gray whale (Eschrichtius robustus)

Pygmy right whale (Caperea marginata)

Humpback whale (Megaptera novaeangliae)

Blue whale (Balaenoptera musculus)

Fin whale (B. physalus)

Sei whale (B. borealis)

Common minke whale (B. acutorostrata)

Antarctic minke whale (B. bonaerensis)

Bryde's whale (B. edeni)*

Pygmy Bryde's whale (B. sp.)*

North Atlantic right whale (Eubalaena glacialis)

North Pacific right whale (E. japonica)

Southern right whale (E. australis)

Bowhead whale (Balaena mysticetus

Sperm whale (Physeter macrocephalus)

*Nomenclature not yet settled for these two species.

These are all considered to be "good" species, i.e., there is no gene flow between any of them, and there is at least some gene flow among populations of each of them.

The task of discriminating populations and establishing boundaries between them is greatly complicated by the fact that most whales migrate. Perrin presented a summary review of migratory patterns in a selected set of the 15 species and noted, when appropriate, how they accord with the archetypes described above He reviewed the gray whale, humpback whale, common minke whale, antarctic minke whale and sperm whale.

Gray whale

The gray whale formerly inhabited both the North Atlantic and North Pacific. It was extirpated from the Atlantic by whaling in medieval to U.S. colonial times. Two remaining populations inhabit the eastern and western sides of the North Pacific. The eastern population was greatly depleted in the 19th Century but has recovered to something like its original abundance. It winters in southern Baja California and migrates to feeding grounds in the Bering and Chukchi Sea. The western population contains only perhaps approximately 100 whales that feed at Sakhalin Island in Russia and winter at unknown locations likely on the coast of southern China near Hainan Island. The two populations thus exemplify Archetype II (two populations with separate feeding and breeding grounds for each population). A possible exception is that some whales in the eastern population may consistently over-summer in smaller feeding grounds in Alaska, Canada and the NW U.S, thus constituting multiple feeding stocks utilizing a single breeding ground (Archetype V).

Humpback whale

The humpback whale occurs worldwide, with oceanic migrations between high-latitude feeding concentrations in coastal regions and lower-latitude wintering areas.

In the North Atlantic, humpback movements exhibit an Archetype V pattern. Whales breeding in the Indies feed in several grounds across the North Atlantic extending from the Gulf of Maine to Iceland and Norway. A deduced second breeding ground off West Africa may contribute whales to feeding grounds in Iceland and Norway. The Caribbean region may actually comprise two separate breeding grounds, with whales migrating through one on the way to the other.

North Pacific humpbacks display a much more complex pattern. At least four major breeding grounds are recognized (Mexico, Costa Rica, Hawaii, and Japan/Philippines. At least some of these may actually consist of adjacent but separate breeding grounds. For example, there have been no matches between individually-identified whales between the Mexican mainland/Baja areas and the offshore Revillagigedo Islands. More than a dozen feeding areas have been identified across the top of the North Pacific, but the boundaries and extents of these are poorly known, and some whales from nearly all the breeding areas have been seen at nearly all the feeding areas. However, some major routes are evident. Most of the whales in California come from Mexico. Nearly all the whales in Alaska come from Hawaii. But there are very many whales in Hawaii that have never been seen in Alaska and must be going somewhere else, perhaps to the western Aleutians or Russia.

In the Southern Hemisphere, the IWC split up the high-latitude feeding grounds around the boundary of Antarctica into six statistical divisions that were thought to roughly correspond to the poorly known breeding grounds at lower latitudes. Since then, some of the supposed breeding populations have proved to be composed of two or more, e.g., Queensland, New Caledonia and Tonga within the putative Eastern Australia breeding population. Similarly, the boundaries between the Southern Ocean areas don't line up well with the breeding areas, with mixing and overlapping populations. The basic applicable Archetype is II but with numerous exceptions and complications.

Most pre-20th Century exploitation of humpback whales was on the breeding grounds. It has been the more recent

industrial whaling on the feeding grounds that has greatly depleted all the stocks.

Common minke whale

The common minke whale occurs throughout the northern hemisphere. A dwarf form (un-named subspecies) also occurs in the Antarctic, seasonally sympatric with the Antarctic species.

In the North Atlantic, the IWC recognizes four feeding stocks: Canadian eastern coast, West Greenland, Central and Northeast. Breeding grounds for the western side are presumably in the SE US, West Indies, and/or northeastern South America and for the eastern side in Iberia, Mediterranean and/or West Africa. Breeding areas for the West Greenland and Central stocks are unknown; they may be in the mid-tropical North Atlantic or in the western or eastern warm-water areas. The migration pattern is further complicated by age and sex segregation, with juveniles migrating first and not as far, and females moving the farthest north, to the ice edge.

Exploitation has occurred mainly in the 20th Century, on both sides of the Atlantic. At least one population is considered depleted.

The pattern in the North Pacific where known is much like that in the Atlantic, with sex and age segregation and possible crisscrossing of migratory routes. However, even less is known about breeding grounds, in terms of both location and numbers. The only fairly clear picture for any stock is that for the whales that winter in the South China Sea and summer in the southern Okhotsk Sea and northern Japanese coast. Practically nothing is known of stock structure in the offshore western Pacific, central Pacific and eastern Pacific.

Antarctic minke whale

The picture for the Antarctic minke whale much resembles that for the humpbacks of the southern hemisphere, but with much less known about migratory connections between the six statistical divisions and the as-yet largely unknown breeding grounds at lower latitudes. Structure on the feeding grounds as indicated by genetic analyses is highly variable, area to area, season to season and year to year. The pattern corresponds to some intermediate or combination of II, III and IV, and perhaps involving type V and VII complexity.

Antarctic minke whales were not exploited significantly until after depletion of the southern stocks of the larger whales, and the takes are not yet thought to have materially affected the stocks.

Bowhead whale

The bowhead is an Arctic whale. It is currently exploited by the U.S., Russia and Canada. Five stocks are recognized by the IWC: Okhotsk Sea, Bering-Beaufort-Chukchi Sea (referred below to Bering Sea Stock), Hudson Bay, Davis Strait and Spitsbergen stocks. The largest stock and the one currently exploited is the Bering Sea stock. It moves through the Bering Strait in spring, east into the Beaufort Sea, back into the Chukchi Sea in summer and down the western side of Kamchatka on the way to its wintering ground in the Bering Sea. The migratory pattern for all five stocks accords with the simple Archetype II. However, some whales have been seen in the Chukchi Sea in spring and early summer when the stock is supposedly in the Beaufort Sea, and it is possible that these comprise a separate feeding stock (again Archetype V).

Bryde's whales

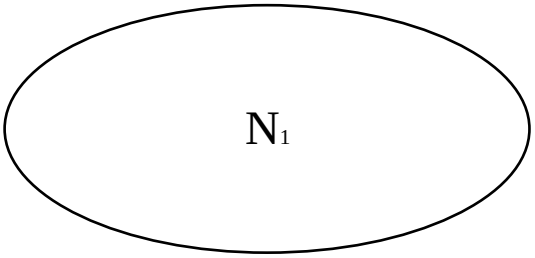
Bryde's whales are tropical/warm temperate whales that do not migrate to high latitudes like the other balaenopterids. There are two species, the pygmy and the ordinary Bryde's whales; only the latter is addressed here. The ordinary Bryde's whale occurs around the world in a band from roughly 40 degrees north to 40 degrees south. Within this species, a smaller coastal form has been recorded from several regions. There are a number of areas of known concentration of the larger offshore pelagic form of the ordinary Bryde's whale, in the Atlantic, Pacific and Indian Oceans. The present IWC stock boundaries in some cases do not accord with these concentrations. Some of the IWC stocks are thought to be depleted; others may be in close-to-original condition or are unassessed. In an alternative

hypothesis of stock structure, northern and southern populations straddle the equator seasonally. Thus the IWC stock areas may be missing part of the populations at one time of the year and contain parts of two populations at another time of the year.

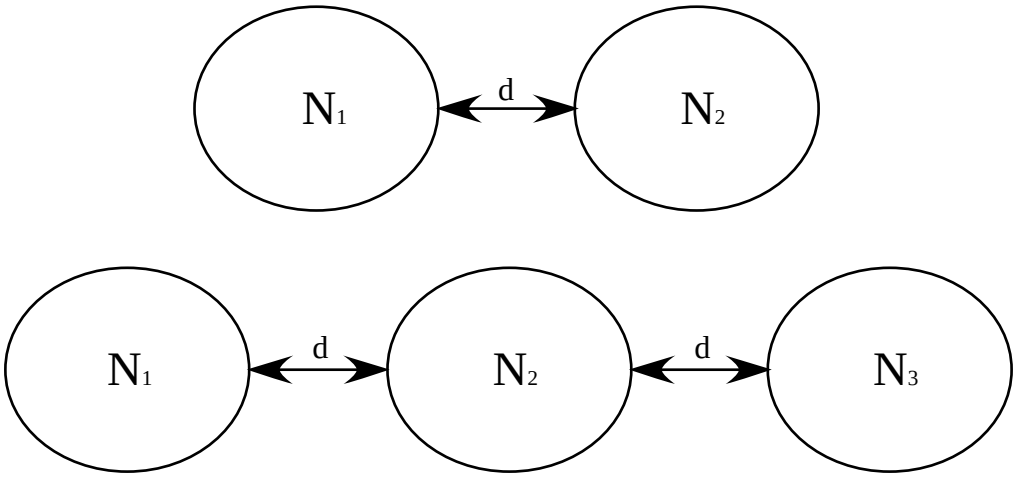
Sperm whale

The sperm whale occurs from the Arctic to the Antarctic in deep water. Very little is known about population structure, because of a complex social system and segregation by age and sex. Long-term associations of groups of females are known, but these are not necessarily closely related. Large males tend to be solitary and migrate to higher latitudes than females. There is a high degree of genetic homogeneity within ocean basins; the reason for this is unknown. There may be reciprocal use of equatorial regions by northern and southern populations, as in the Bryde's whale. Under one hypothesis of stock structure in the North Pacific, at different times of the year, differing populations and different segments of populations can be found in the same area. However, tag returns from juvenile males indicate extensive movements across the putative stock boundaries during the 25+ years before sexual maturity.

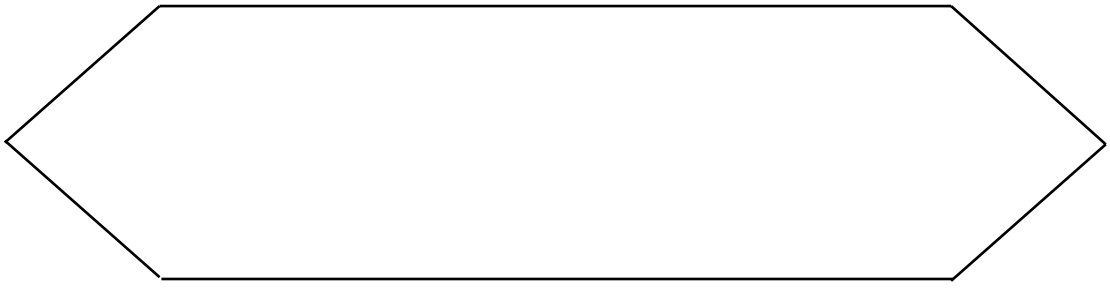
The status of sperm whale stocks is unknown. They were heavily exploited in open-boat whaling in the 19th century and again in some areas more recently during modern industrialized whaling. The obscure population structure and unreliability of much of the recent catch data have rendered assessment efforts ineffectual.



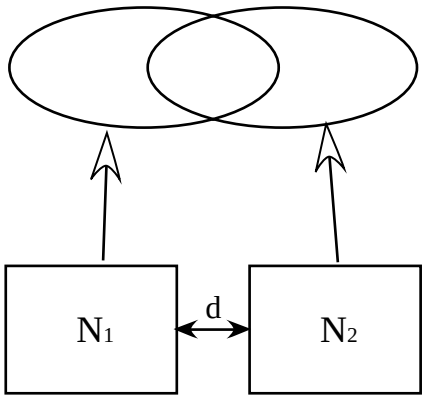
Archetype I. Panmixia



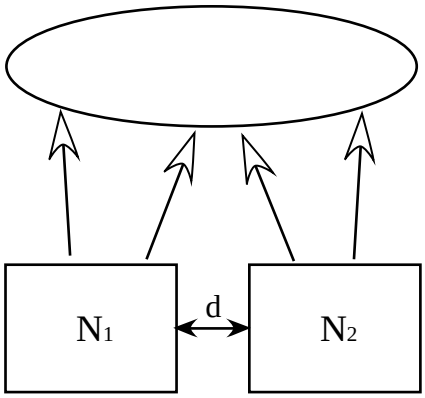
Archetype II. Stepping-stone. There can be either two or three populations. Dispersal occurs only between adjacent populations.



Archetype III. Diffusion-type isolation-by-distance.

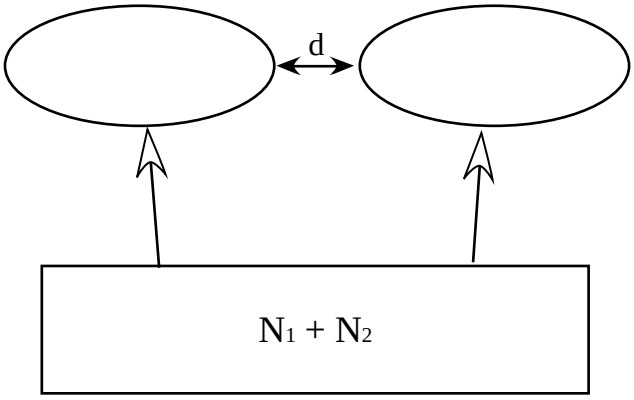


Overlap of feeding grounds equals 50%



Overlap of feeding grounds equals 100%

Archetype IV. Two discrete breeding grounds with feeding grounds that overlap partially or completely. Ovals indicate feeding grounds while rectangles depict breeding grounds. Open ended arrows indicate migratory routes while closed arrows indicate dispersal.



Archetype V. A single breeding stock with two separate feeding grounds. Animals follow their mothers to the feeding ground and exhibit strong feeding ground fidelity. Ovals indicate feeding grounds while rectangles depict breeding grounds. Open ended arrows indicate migratory routes while closed arrows indicate dispersal due to females occasionally changing feeding grounds.